International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June 2018 www.irjet.net p-ISSN: 2395-0072

# FAKE PRODUCT REVIEW MONITORING

**Madhura N Hegde [1], Sanjeetha K Shetty [2], Sheikh Mohammed Anas [3], Varun K [4]**

[1]*Assistant Professor, Dept. of Information Science Engineering, Sahyadri College of Engineering and Management, Karnataka, India*

[2,3,4]*Student, Dept. of Information Science Engineering, Sahyadri College of Engineering and Management, Karnataka, India*

---

**Abstract -** Online product review on shopping experience in social media has promoted users to provide customer feedback. Nowadays many e-commerce sites allow customers to write their opinion on the product which they buy in the form of reviews or ratings. The reviews given by the customer can build or shatter the good name of the product. Due to this reason company personnel gets an idea of standing's of their product in the market. In order to demote or promote the product, spiteful reviews or fake reviews, which are deceptive, are posted in the ecommerce site. This result will lead to potential financial losses or larger amount of growth in business. We propose a project which focuses on detecting fake and spam reviews by using sentiment analysis and removes out the reviews which contains vulgar and curse words and make the e-commerce site fake review free online shopping center.

***Key Words*: Sentiment analysis, Content similarity, Review deviation.**

## 1. INTRODUCTION

### 1.1 Data Mining

Data mining is a process of thoroughly scrutinizing huge store of data to discover patterns and trends that go beyond simple analysis. Data mining uses worldly-wise mathematical algorithms to fragment the data and evaluate possibility of future event. Opinion mining is a kind of Natural Language Processing for tracking the attitude of public about a particular product. A fundamental task in opinion mining is classifying the polarity of a given text whether the expressed opinion in a document or a sentence is positive, negative, or neutral. The birth of Social Media has brought about enormous influence in all spheres of life. From casual comments in Social Networking sites to more important studies regarding market trends and analyzing strategies, all which have been possible largely due to rising impact of Social Media data. A predominant piece of statistics for most of us during the decision-making process is "what other people think". One such form of Social Media data are Product reviews written by customers reporting their level of contentment with a specific product. These reviews are helpful for both individuals and business firms. These review systems motivate some people to enter their fake review to promote some products or downgrade some others. The main reason for this action is to make more profit by writing unfaithful reviews and false ratings .So in order to make products and services trustable, these fake opinions must be detected and removed. Detection techniques are used to discover fake reviews. The System is developed using three stages- Preprocessor, Fake review detector and Classifier. Web scraper is used to extract required data from the website. Preprocessor is used to filter out abusive reviews and transform the reviews to a proper format. Fake review detector detects the fake reviews using sentiment computing, review deviation and content similarity techniques and thus reviews are marked fake and genuine. Finally, the classifier takes labeled and unlabeled samples as input and labels the unlabeled samples.

## 2. ARCHITECTURE OF THE PROPOSED MODEL

Architecture design gives the real world view of the system. **Fig-1** represents the architecture design of Fake Product Review Monitoring. The reviews which are part of online data on websites are scraped by the web crawler. This data is taken for preprocessing where the data is transformed into required format and the abusive reviews are removed. Fake reviews are next identified by Fake Review Detector. This is taken as training data for the classifier which classifies fake and genuine reviews.

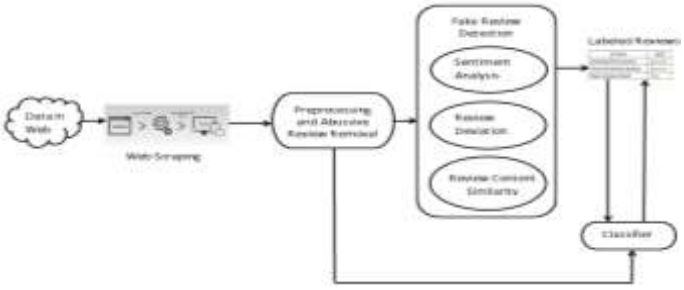

**Fig -1.**: Architecture diagram

Data Flow Diagram represents the flow of information or data in the system. The square boxes represent the entities, ovals represent the process and named arrows represent the direction of flow of information. Figure shows the diagram of Fake Review Detector. **Fig-2** shows the data flow diagram of Fake Review Detector. It has three entities Fake Review Detector, FRD WebCrawler and Classifier and three

processes namely WebScraping, Preprocessing and FakeReviewDetection.

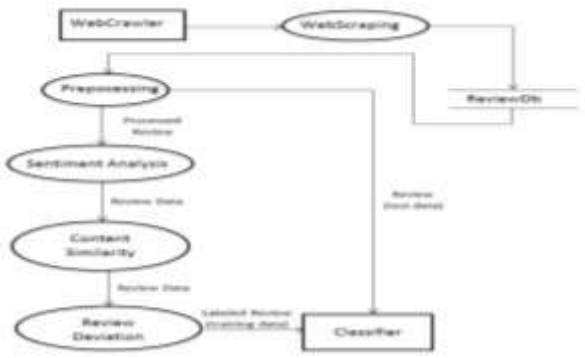


**Fig -2.:** Data flow diagram

## 3. METHODOLOGY

### 3.1 Sentiment Analysis

Sentiment analysis is used to understand writers emotion. We define three list of words; positive vocabulary, negative vocabulary and neutral vocabulary, which consists of positive, negative and neutral words. Every review is passed to nltk classifier which calculates the sentiment score of the reviews. The rating is normalized using the equation $NR_i = R_i - 3 - (R_i - 3)/2$ . The difference between normalized rating and sentimen score is found. If the calculated difference is greater than 0.5 the review is considered to be fake.

### Psuedo Code for Sentiment Analysis

1: classifierNaiveBayesClassifier.train(train_ set)

2: i <- 0

3: For r1 in review:

4: neg <- 0

5: pos <- 0

6: words <- r1.split(' ')

7: For word in words:

8: classResult <- classifier.classify( word feats(word))

9: If classResult == 'neg' :

10: neg = neg <- 1

11: End If

12: if classResult == 'pos' :

13: pos = pos <- 1

14: End If

15: End For

16: positive <- float(pos) / len(words)

17: negative <- float(neg) / len(words)

18: score <- positive - negative

19: nrating[i] <- rating[i] - 3 - (float(rating[i])-3) / 2

20: If | nrating[i]-score| >0.5

21: label[i] ¡- "fake"

22: End If

23: i <- i + 1

24: End For

### 3.2 Content Similarity

Content similarity is performed on the reviews given by same user. We use cosine similarity to obtain similarity of two reviews. If the cosine value is greater than 0.5 the review is considered to be fake.

### Psuedo Code for Content Similarity

1: j <- 0

2: For r1 in review:

3: i <- 0

4: For r2 in review:

5: text1 <- r1

6: text2 <- r2

7: If ( j <i and (author[i]==author[j])) :

8: If ( j <i and (author[i]==author[j])) :

9: vector1 <- text to vector(text1)

10: vector2 <- text to vector(text2)

11: cosine <- get cosine(vector1, vector2)

12: If cosine >= 0.5 :

13: label[i] <- "fake"

14: label[j] <- "fake"

15: End If

16: i <- i+1

17: End For

18: j <- j+1

19: End For

### 3.3 Review Deviation

Review deviation is based on the theory that if all customers give less rating and one customer gives high rating or vise versa, then that rating/review is considered to be fake. First we calculate average of all the ratings and then we check to what extent it deviates from the origin. If the deviation is

greater than or equal to 2, then the review is considered to be fake.

### Psuedo Code for Content Similarity

1: sum <- 0

2: n <- 0

3: For r1 in rating:

4: sum <- sum + r1

5: n <- n + 1

6: End For

7: avg <- sum/n

8: For r1 in rating: 9: d <- | r1-avg | 10: If ( d >= 2) :

11: label[i] <- “fake”

12: End If

13: i <- i + 1

14: End For

## 5. RESULTS

The system developed can detect 111 fake reviews out of 300 reviews. The classifier built detected 18 fake reviews out of 52 reviews. The Fig -3 shows graph for the time taken by the system to process different number of reviews.

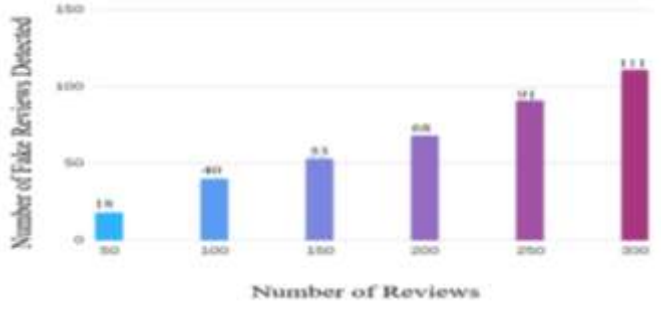


**Fig -3.:** Bar graph: Number of reviews taken v/s Number of fake reviews detected

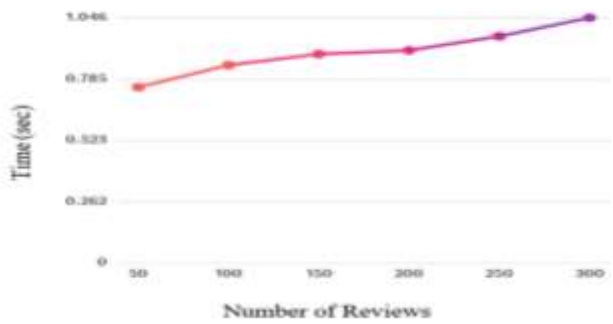


**Fig -4.:** Bar graph: Performance of the system

## 4. CONCLUSIONS

This work has investigated the user's request and has given the particular prediction and safety measure for that input. The K-Means clustering technique was used to find the clusters and fatality for the flight crash investigation. This method will also consider other factors like efficiency, weather impact and schedules of other aircraft. Results showed that the proposed algorithm obtained prediction based on user's input with efficiency. The proposed algorithm also provided improved search results for the query given by the user. Possible future work is to improve the efficiency and also increase the count of clusters used.

## ACKNOWLEDGEMENT

It is with great satisfaction and euphoria that we are submitting the paper on “Fake Product Review Monitoring". We are profoundly indebted to our guide, Mrs. Madhura N. Hegde, Assistant Professor, Department of Information Science & Engineering, for innumerable acts of timely advice, encouragement and we sincerely express our gratitude. We also thank her for her constant encouragement and support extended throughout.
Finally, yet importantly, we express our heartfelt thanks to our family & friends for their wishes and encouragement throughout our work.

## REFERENCES

[1] Shashank Kumar Chauhan, Anupam Goel, Prafull Goel, Avishkar Chauhan and Mahendra K Gurve, “Research on product review analysis and spam review detection”, 4th International Conference on Signal Processing and Integrated Networks(SPIN) 2017, ISBN (e):978-1-5090-2797-2, September-2017, pp. 1104-1109.

[2] Sonu Liza Christopher and H A Rahulnath, “Review authenticity verification using supervised learning and reviewer personality traits”, International Conference on Emerging Technological Trends (ICETT) 2016, ISBN (e): 978-1-5090-3752-0, pp. 1-7.

[3] Ruxi Yin, Hanshi Wang and Lizhen Liu, "Research of integrated algorithm establishment of spam detection system", 4th International Conference on Computer Science and Network Technology (ICCSNT) 2015, ISBN (e): 978-1-4673-8173-4, pp. 390-393 .

[4] Xinkai Yang, "One methodology for spam review detection based on review coherence metrics", International Conference on Intelligent Computing and Internet of Things 2015, ISBN (e): 978-1-4799-7534-1, pp. 99-102.

[5] Hamzah Al Najada; Xingquan Zhu, "iSRD: Spam review detection with imbalanced data distributions", Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014), ISBN (e): 978-1-4799-5880-1.\

[6] SP.Rajamohana, Dr.K.Umamaheshwari, M.Dharani, R.Vedackshya, "A survey on online review SPAM detection techniques", International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT) 2017, ISBN(e): 978-1-5090-5778-8. ", International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT) 2017, ISBN(e): 978-1-5090-5778-8.

## BIOGRAPHIES

Madhura N. Hegde
Assistant Professor,
Areas of Interest: Data mining, Image processing.
Sahyadri College of Engineering and Management

Sanjeetha K Shetty
Student
Sahyadri College of Engineering and Management

Sheikh Mohammed Anas
Student
Sahyadri College of Engineering and Management

Varun K
Student
Sahyadri College of Engineering and Management